DE GRUYTER MOUTON
DOI 10.1515/cog-2013-0022 — Cognitive Linguistics 2013; 24(4): 633–665

Ewa Dąbrowska

Functional constraints, usage, and mental grammars: A study of speakers' intuitions about questions with long-distance dependencies

Abstract: This paper describes an experimental study which attempts to reconcile two usage based approaches to questions with long distance dependencies (LDDs): the Lexical Template Hypothesis (Dąbrowska 2004, 2008; Verhagen 2005, 2006) and Goldberg's BCI ("Backgrounded Constituents are Islands") constraint (Goldberg 2006; Ambridge and Goldberg 2008). The study replicates Ambridge and Goldberg's (2008) results supporting the BCI constraint; but it also shows that (1) LDD questions with *think* and *say*, the verbs which are part of the hypothesised templates, are judged to be more acceptable than predicted by BCI and (2) BCI cannot explain complementizer effects (why LDD questions with *that* are judged less acceptable than questions without *that*). The results also suggest that there are considerable individual differences in speakers' sensitivity to the constraint.

Thus, the two hypotheses are complementary: BCI explains why certain LDD questions are more acceptable than others, and hence accounts for differences in the frequency of prototypical and unprototypical LDD questions, while the lexical template hypothesis explains the effects of the frequency of use on speakers' mental grammars.

Keywords: functional explanation, long-distance dependencies, usage-based models, individual differences, frequency

Ewa Dąbrowska: Northumbria University, Newcastle, UK. E-mail: Ewa.Dabrowska@unn.ac.uk

1 Introduction

Questions and other constructions with long distance dependencies (henceforth LDDs) have played an important role in the development of syntactic theory, especially in the generative framework. More recently, they have also attracted the attention of cognitive linguists (see e.g. Ambridge and Goldberg 2008; Dąbrowska 2004, 2008; Goldberg 2006; Verhagen 2005, 2006). These structures are interest-

ing because they exhibit a dependency between a filler in the main clause and a gap in a subordinate clause, as in example (1). Such dependencies are often called 'unbounded', as, in principle, there can be any number of clauses intervening between the filler and the gap (cf. 2).

(1) What did John claim that I did __?

(2) What did John claim that Tom thought that Claire imagined that I did __?

However, questions with dependencies spanning more than one clause boundary are virtually nonexistent in real life: the British National Corpus (2001), for example, which consists of over 100 million words, does not contain a single instance. Moreover, attested LDD questions are extremely stereotypical: the main clause nearly always contains just the question word followed by the auxiliary *do*, the pronoun *you*, and the verb *think* or *say*. In fact, 67% of the LDD questions in the BNC have the form *WH do you think S-GAP?* or *WH did you say S-GAP?*, where S-GAP is a subordinate clause with a missing constituent. Most of the remaining questions are minimal variations on these patterns: that is to say, the main clause contains a different subject *or* a different verb *or* a different auxiliary *or* an additional element like an adverbial or complementizer; only 4% depart from the prototype in more than one respect. LDD questions in child-directed speech are even more stereotypical, with the two formulas accounting for 94% of all instances (Dąbrowska et al. 2009).

This has led some researchers working in the usage-based framework (Dąbrowska 2004, 2008; Verhagen 2005, 2006) to suggest that speakers store lexically specific templates corresponding to the frequently occurring combinations, with slots for the variable elements. According to this view, speakers produce "prototypical" LDD questions, i.e. those that match one of the hypothesised templates, simply by inserting lexical material into the slots. For instance, a question like *What do you think he will buy?* can be produced by inserting *what* into the first slot and *he will buy (WH-GAP)* into the second slot in the lexical formula *WH do you think S-GAP?* To produce nonprototypical questions, speakers need to modify the template as well as insert lexical material into the slots. This could be accomplished by relying on proportional analogy, as illustrated in (3); note that the parts in capitals correspond to meaning while italics represent phonological form.

(3) YOU THINK HE WILL BUY STH: WHAT? is to *What do you think he will buy?*
as
SHE HOPES HE WILL BUY STH: WHAT? is to ???

The Lexical Template Hypothesis (LTH) makes several testable predictions: prototypical LDD questions should be produced more fluently, remembered better, judged to be more acceptable, and acquired earlier by children. All of these predictions have been confirmed (Dąbrowska 2008; Dąbrowska et al. 2009).

Goldberg (2006) addresses a different aspect of LDD constructions, namely, the well-known fact that some syntactic constituents are islands: long distance dependencies cannot reach into complex NPs (4), sentential subjects (5), complements of manner of speaking and factive verbs (6)–(7), or presupposed adjuncts (8). (All examples are from Goldberg 2006.)

(4) *Who did she see that the report was about __?

(5) *Who did that she knew __ bother him?

(6) ??Who did she whisper that he left __?

(7) ??Who did she know that he left __?

(8) ??What did she leave the movie 'cause they were eating?

Generative linguists account for the ungrammaticality of such sentences by appealing to a syntactic constraint (subjacency). Goldberg proposes a different explanation, based on information structure, which she calls BCI ("backgrounded constituents are islands"). She points out that the WH word is always focussed, while complements of factive and manner of speaking verbs, complex NPs, sentential subjects and presupposed adjuncts are all backgrounded. Since the WH word and the gap refer to the same participant, and the same participant cannot be both focussed and backgrounded at the same time, sentences like (4)–(8) involve a clash between the information-structure properties of the LDD construction and the other constructions involved, and are therefore unacceptable.

Goldberg (2006) argues that the BCI account is better motivated than subjacency and explains a wider range of facts. Ambridge and Goldberg (2008) describe an experimental study which provides further support for the proposed principle. The experiment consisted of two parts: an acceptability judgment task and a 'negation test'. In the acceptability judgment task, participants were asked to rate the acceptability of LDD questions and the corresponding declaratives, e.g.

(9) What did Jess think that Dan liked?

(10) Jess thought that Dan liked the cake.

In the negation test, the participants' task was to judge to what extent a sentence with a negated complement-taking verb, such as (11), implies the truth of the negation of the subordinate clause, e.g. (12):

(11) Maria didn't think that Ian liked the cake.

(12) Ian didn't like the cake.

The negation test exploits the fact that presupposed constituents are always backgrounded, and that presupposition survives under negation. The test measures the extent to which speakers judge the information in the subordinate clause to be presupposed, and thus backgrounded. Ambridge and Goldberg manipulated the verb in the main clause, using three types of verbs which differ in the degree to which they presuppose the truth of the subordinate clause: factives (*realize*, *remember*, *notice*, *know*), manner-of-speaking verbs (*whisper*, *stammer*, *mumble*, *mutter*), and "bridge verbs" (*say*, *decide*, *think*, *believe*). BCI predicts that there should be a correlation between responses to the negation test and the acceptability of LDD questions with a given verb. To control for lexical effects, what Ambridge and Goldberg actually looked at was the correlation between the participants' responses on the negation task and their "difference scores", i.e. the rating for the declarative sentence minus the rating for the corresponding interrogative. Since individual data are noisy, they computed the correlation between the average difference score and the average negation test result for each verb. Their results indeed revealed a strong negative correlation ($r = -0.83$, $p = 0.001$) between responses on the negation test and difference scores. Thus the prediction was confirmed.

BCI is a general functional constraint which explains why certain combinations of words do not occur. It is not necessarily incompatible with the Lexical Template Hypothesis (LTH). One could argue that functional constraints do not shape speakers' mental grammars directly: they shape usage, which in turn shapes grammars. Thus BCI and LTH could in principle describe different aspects of the same phenomenon. However, Ambridge and Goldberg argue against the lexical template, or "item-based", account, pointing out that the acceptability ratings for questions with *think* and *say* are not significantly higher than predicted by BCI. This can be seen in Figure 1, which shows the average difference scores for the 12 verbs used in the study plotted against their average negation test values, the regression line computed on the basis of these figures, and 95% confidence intervals for individual values. As we can see from the figure, *think* is almost on the regression line (i.e., its actual value is very close to the predicted value). For *say*, the actual value is 1.62 standard deviations below the predicted

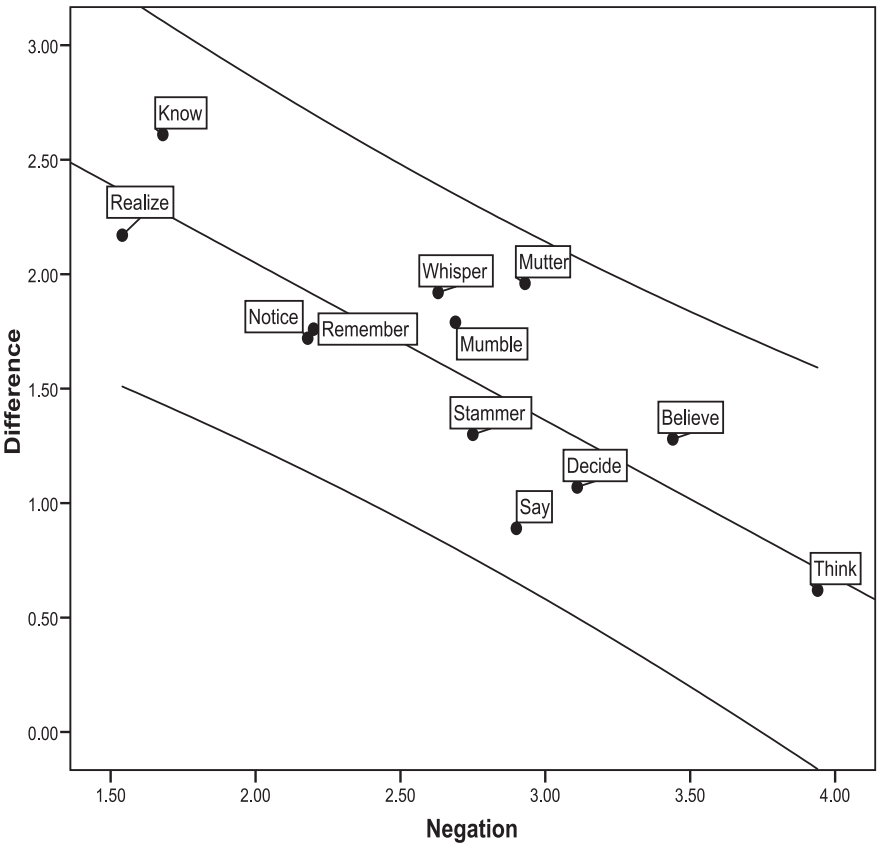

Fig. 1: Relationship between negation test scores and difference scores (dispreference for question scores) (redrawn from Ambridge and Goldberg 2008)

value, and hence it also does not reach outlier status, the usual criterion for outliers being 1.96 standard deviations from the mean.

The BCI account has much to recommend it. In principle, it could account for a wider range of phenomena than the LTH – not just LDD questions, but also related constructions. Furthermore, it provides a more satisfactory account in that it explains why there are differences in frequency in the first place: verbs whose meanings are most compatible with the meaning of the construction are likely to be used in it most often. However, it has some problems of its own. BCI provides a general pragmatic explanation of why certain sentences with LDDs are acceptable while others are not which one would expect to be valid for other languages as well. However, not all languages are like English in this respect:

some are more permissive, allowing extraction out of complex NPs as well as verb complement clauses, while others are more restrictive and don't allow extraction out of either (see Hawkins 2004). BCI alone cannot explain these crosslinguistic differences. Secondly, even within the same language, there are differences in native speakers' intuitions which may be related to linguistic experience. For instance, English-speaking linguists tend to accept a wider range of LDD questions than non-linguists; and generative linguists are more tolerant of complex NP violations than other linguists and naïve speakers (Dąbrowska 2010). Last but not least, while the findings reported by Ambridge and Goldberg explain the differences in acceptability due to using different verbs, Dąbrowska (2008) demonstrated other types of prototypicality effects as well. Specifically, adding an overt complementizer or an additional complement clause, or using an auxiliary other than *do* has a different effect on LDD questions than on the corresponding declaratives. It remains to be seen whether BCI can explain these findings as well.

It should also be pointed out that the Ambridge and Goldberg study was designed to test the BCI hypothesis, and the method they employ is not ideal for testing the LTH, for three reasons. First, Ambridge and Goldberg used data for all 12 verbs as input for the regression equation. This is reasonable when testing BCI; however, incorporating data for *think* and *say* into the regression model makes it less likely that these verbs will turn out to be outliers. A fairer test for the LTH would compute the regression equation on the basis of the other verbs, and then see how accurately it predicts participants' acceptability judgments for questions with *think* and *say*, given the negation test ratings for these verbs. To see that this matters, compare Figure 1 above (where the regression line was computed on the basis of the entire data set) with Figure 2 (in which the regression line was computed on the basis of all verbs except *think* and *say*). The difference between the predicted values and the actual values is now much larger, and *say* is now quite close to the lower 95% confidence interval – in spite of the fact that with only 10 data points, the confidence intervals are quite wide.

Secondly, the use of difference scores may have masked interactions between construction type and acceptability. The rationale for using difference scores was to control for lexical effects of the verb. Acceptability ratings for an LDD question with a particular verb depend not just on the compatibility of the verb with the LDD construction, but also on other properties of the verb: its frequency, stylistic markedness, compatibility with the complementation construction, etc. Subtracting the LDD question rating for a particular verb from the rating for the corresponding declarative, Ambridge and Goldberg argue, allows us to eliminate the effects of these irrelevant factors, providing a more accurate measure of the verb's "dispreference for extraction". However, the use of difference scores as-

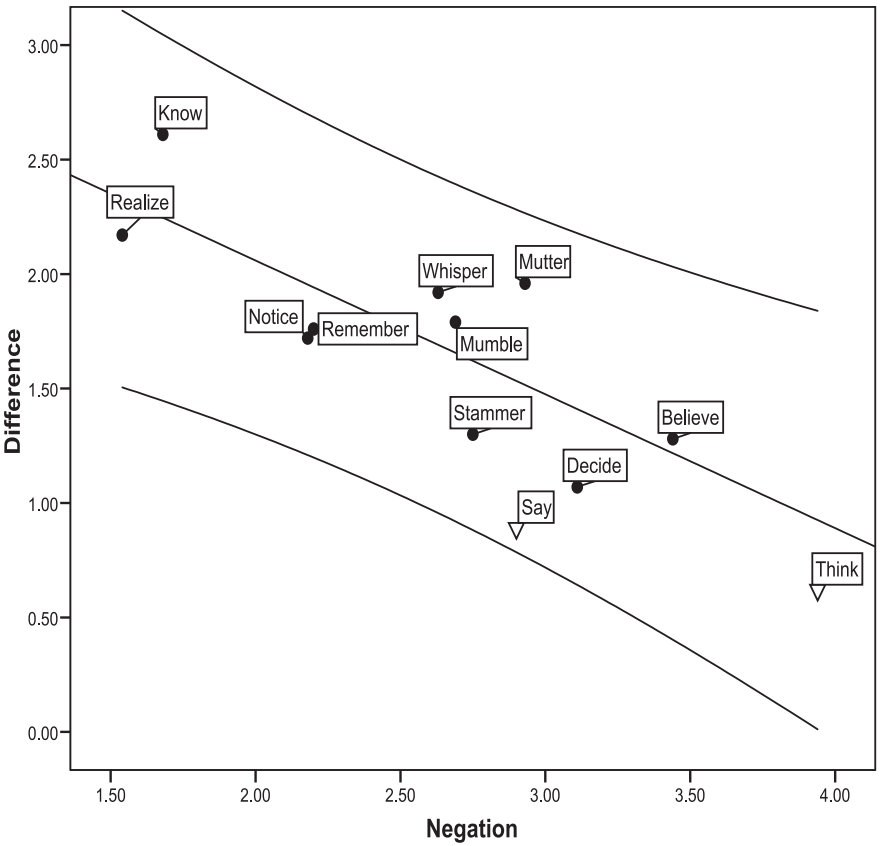

Fig. 2: Relationship between negation test scores and difference scores, computed on the basis of all verbs except *think* and *say* (drawn from data provided in Ambridge and Goldberg)

sumes that lexical effects are the same regardless of construction – and this may not be the case. A low difference score may mean that a verb works particularly well in an LDD construction *or* that it doesn't work particularly well in the corresponding declarative. In other words, using difference scores is a legitimate move when testing BCI, but not when testing LHT: by subtracting question ratings from declarative ratings, we may be subtracting away the very effects that we are trying to find.

To see that there is some substance to this argument, compare Figure 2 and Figure 3, which shows the relationship between the negation test and acceptability ratings for questions. There is still a significant relationship, although the cor-

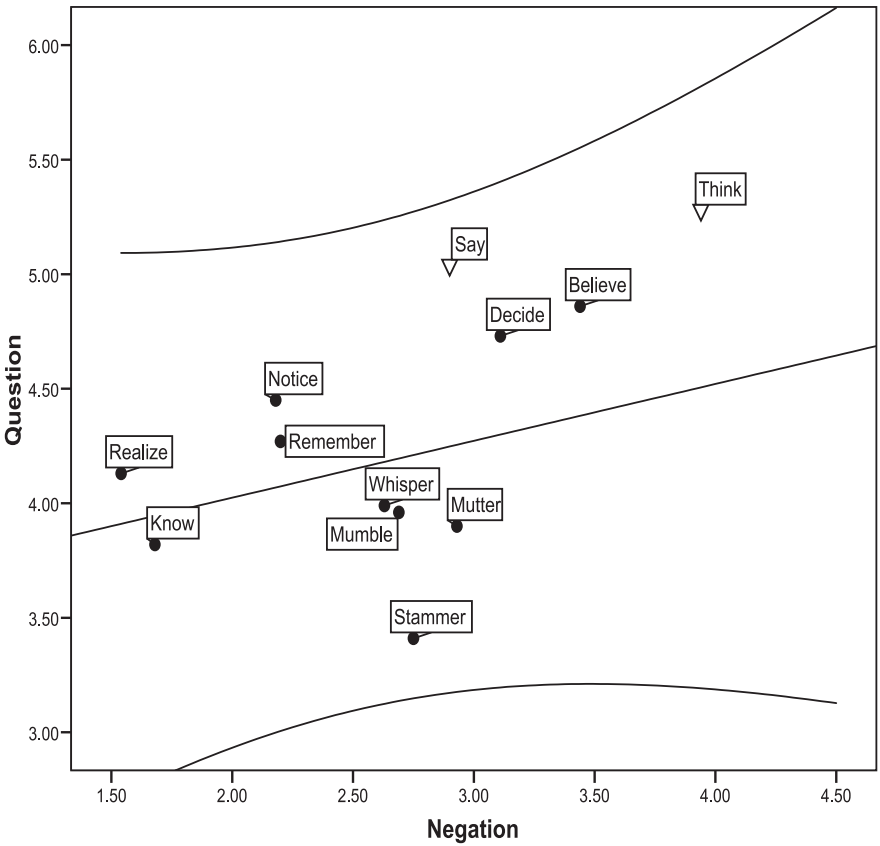

Fig. 3: Relationship between negation test scores and acceptability ratings for LDD questions, computed on the basis of all verbs except *think* and *say* (drawn from data provided in Ambridge and Goldberg)

relation coefficient is lower (r = 0.58, p = 0.047). This shows that the argument for BCI does *not* depend on the use of difference scores. (The correlation now has a positive coefficient because we are looking at the relationship between raw scores: the higher the negation score, the higher the acceptability rating for LDD questions. Using difference scores reverses the relationship: the higher the negation score, the lower the difference between declaratives and LDD questions.) Crucially, however, *think* has moved even further away from the regression line, although it is still within 95% confidence intervals for individual values.

To fall above the upper confidence interval, questions with *think* would have to receive ratings of 5.8 or above, and questions with *say*, 5.3 or above. Given the

way the stimuli used in the experiment were constructed, this is extremely unlikely – in fact, virtually impossible in the case of *think*. LDD questions are difficult to produce and understand because the filler must be held in memory while the rest of the sentence is being processed (see Frazier and Clifton 1989; Kluender and Kutas 1993; Hawkins 1999), and consequently they are usually given lower acceptability ratings than the corresponding declaratives. (Note that all the difference scores in Ambridge and Goldberg are positive, i.e. the averaged ratings for questions were always lower than those for the corresponding declaratives.) The average ratings for declaratives with *think* and *say* was 5.9. The ratings for questions with these verbs should be lower than this – considerably lower, in fact, since the questions used in the acceptability judgment tasks departed from the hypothesised formula in two respects: they contained an overt complementizer and a lexical rather than pronominal subject. Moreover, the auxiliary was always in the past tense (e.g. *What **did** Jess think that Dan liked?*); however, with verbs like *think*, which designate permanent states, the present tense is much more frequent.

To summarize: the Ambridge and Goldberg study was designed to test BCI, not the LTH. Because of the properties of the stimuli used in the experiment and the way the data were analysed, it was extremely unlikely to reveal prototypicality effects even if they do exist. Given that there is independent evidence for lexical templates for LDD questions and that the two accounts are not necessarily incompatible, it seems appropriate to address this issue again. The study described in this paper is a partial replication and extension of the Ambridge and Goldberg experiment which uses different stimuli sentences, designed to offer a fairer test of the Lexical Template Hypothesis. The study is intended to answer three questions:

1. Can the strong relationship between backgrounding and acceptability of LDD questions found by Ambridge and Goldberg be replicated with these different stimuli?
2. Can the particularly high ratings that participants give for LDD questions with *think* and *say* be explained by BCI, or are such questions more acceptable than one would expect on the basis of the semantic properties of these verbs (as the lexical template account predicts)?
3. Can BCI also account for other prototypicality effects in LDD questions, notably the fact they are judged to be more acceptable when they do not contain a complementizer (cf. Dąbrowska 2008)?

2 Method

2.1 Participants

64 undergraduate students enrolled in a first year course in linguistics at the University of Sheffield participated in the study. Since entry on the course is fairly selective, it is likely that the participants had above average IQs and metalinguistic skills. All participants had completed an introductory linguistics course which included five lectures on syntax, but have not had any instruction on LDD constructions. Four of the participants did not complete the entire questionnaire, and hence their data were discarded. Thus the final sample comprised 60 participants.

2.2 Materials

As in the Ambridge and Goldberg study, participants were asked to complete two tasks: an acceptability judgment and a negation task.

Acceptability judgment

In the Ambridge and Goldberg study, the sentences had the form

> *What did NP1 VERB1 that NP2 VERB2?*
> (e.g. *What did Jess think that Dan liked?*)

and

> *NP1 VERB1-PAST that NP2 VERB2 + APPROPRIATE NP*
> (e.g., *Danielle thought that Jason liked the cake.*)

Thus the questions departed from the hypothesized prototype in two respects: they had third person rather than second person subjects and they contained an overt complementizer. Furthermore, the auxiliary was always in the past tense, which is somewhat unusual with verbs like *think*.

In the current study, all questions had second person subjects (*What do you think . . . ?*) while all declaratives had first person subjects (*I think . . .*). Sixteen different main clause verbs were used. Eight of the verbs designated various permanent states (*think*, *understand*, *know*, *suspect*, *see*, *hope*, *believe*, *mean*) and were used in the present tense (*What do you think . . . ?*, *I think . . .*); the other eight (*say*, *imply*, *prove*, *speculate*, *notice*, *affirm*, *swear*, *complain*) designated discrete events and were used in the past tense (*What did you say . . . ?*, *I said . . .*).

There were two versions of the questionnaire. In each version, half the verbs were used with an overt complementizer and half without a complementizer. Verbs which were used with complementizers in version 1 were used without a complementizer in version 2, and vice versa. In both versions, a particular verb occurred either with or without a complementizer in the declarative as well as the interrogative. As in the Ambridge and Goldberg study, different nouns were used in the declarative and the interrogative to disguise the relationship between the two sentences. Examples of experimental sentences are given in Table 1.

The questionnaire also contained 16 ungrammatical fillers: four involving a dependency reaching into a noun-complement clause (e.g. **What did you discover the fact that the criminals stole?*), four *that*-trace violations (**What do you think that got lost?*), four sentences in which agreement in the main clause was marked on the main verb as well as the auxiliary (**His cousin doesn't thinks we lied*), and four containing the negative particle *not* without an auxiliary in the main clause (**Her husband not claimed they left*).

The 48 sentences (16 questions, 16 declaratives, and 16 ungrammatical controls) were arranged in a pseudorandom order with the following constraints:

1. No two sentences of the same type (question, declarative or control) occurred immediately next two each other; and
2. Experimental sentences containing the same main verb were separated by at least three other sentences.

Within each version, four different pseudorandom orders were used.

Negation test

Each item in the negation test consisted of two parts: an entailing sentence and a test sentence. The entailing sentences were derived from the declaratives in part 1 by replacing the main clause subject *I* with a proper name and negating the verb (see Table 1 for examples). The test sentences were constructed by negating the

Table 1: Examples of stimuli used in the experiment

Condition	Example
Acceptabilty judg.	
Interrogative	What did you say (that) Laura eats?
Declarative	I said (that) Ian eats fish.
Negation test	
Entailing sentence	Neil didn't say (that) Ian eats fish.
Test sentence	Ian doesn't eat fish.

subordinate clause in the entailing sentence. As in the acceptability judgment task, half of the entailing sentences contained a complementizer and the other half did not. There were two versions of the negation test corresponding to the two versions of the acceptability judgment test, with a given verb occurring either with or without a complementizer in both tasks. In each version, the items were presented in one of four random orders.

2.3 Procedure

Both tasks were administered via a written questionnaire completed at the end of a lecture. All participants completed the acceptability judgment task first, immediately followed by the negation task. They were given as much time as they needed, but the majority were able to finish both tasks in about 15 minutes.

The instructions given to the participants were the same as those used by Ambridge and Goldberg, with two small modifications. First, participants were asked to respond by writing a number in a blank instead of circling a number on a pre-printed scale. Secondly, in order to prevent errors involving reversal of the scale, a sentence was added at the end of the instructions section reminding participants about the orientation of the scale (1 = completely bad/not true; 7 = completely good/true). One version of the test, including full instructions, is provided in the Appendix.

3 Results and discussion

3.1 General

Averaging across participants and sentences with and without an overt complementizer, the mean acceptability ratings were 5.31 (SD = 0.67) for questions, 6.26 (SD = 0.44) for declaratives, and 2.55 (SD = 0.74) for ungrammatical controls. The mean score for the negation test was 3.62 (SD = 0.78). Acceptability ratings for questions with *think* and *say* were close to ceiling (6.72 and 6.58 respectively). Ratings for questions with the other verbs ranged from 3.52 to 6.17. The descriptive statistics for individual verbs in all experimental conditions are given in Table 2.

Table 2: Descriptive statistics

Verb	Overt complementizer						Zero complementizer					
	Acceptability: Questions		Acceptability: Declaratives		Negation		Acceptability: Questions		Acceptability: Declaratives		Negation	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
believe	5.53	1.655	6.50	1.009	4.10	1.900	6.33	.922	6.67	.922	4.03	1.520
hope	5.53	1.613	6.93	.254	3.70	1.317	6.23	.774	6.90	.403	3.70	1.022
imply	6.03	1.098	6.43	1.073	4.17	1.392	6.30	1.149	6.30	1.149	4.27	1.337
mean	4.87	1.717	5.97	1.351	4.30	1.705	6.03	1.159	6.17	1.289	4.10	1.125
prove	5.03	1.771	6.43	1.006	4.33	1.583	5.33	1.516	6.50	.820	4.40	1.380
say	6.30	1.088	6.80	.484	3.60	1.163	6.87	.434	6.83	.461	4.00	1.414
see	4.70	1.725	6.47	.776	3.43	1.888	3.53	1.697	6.57	1.073	3.43	2.144
speculate	5.30	1.535	6.00	1.313	4.27	1.461	4.83	1.642	4.87	1.697	4.13	1.106
affirm	4.57	1.695	5.70	1.512	4.10	1.296	5.40	1.522	5.80	1.031	4.00	1.414
complain	3.63	1.884	5.23	1.736	2.50	1.456	3.40	1.940	4.43	1.716	2.60	1.673
know	4.53	1.548	6.80	1.095	2.07	1.552	4.90	1.517	6.60	1.003	2.07	1.617
notice	5.07	1.388	6.53	1.432	2.23	1.569	5.27	1.964	6.70	.702	2.27	1.574
suspect	5.67	1.241	6.77	.626	3.03	1.752	6.27	1.285	6.80	.484	4.23	1.906
swear	5.53	1.502	4.93	2.243	4.07	1.311	5.20	1.606	4.73	1.760	4.40	1.380
think	6.53	.681	6.83	.913	4.93	1.388	6.90	.305	6.93	.254	5.07	1.258
understand	4.17	1.763	6.57	1.104	1.90	1.423	4.13	1.224	6.63	.718	2.47	1.961
All verbs	**5.19**	**0.87**	**6.31**	**0.51**	**3.55**	**0.91**	**5.43**	**0.77**	**6.21**	**0.58**	**3.70**	**1.01**

3.2 Testing the BCI: the role of the verb

Figure 4 shows the mean acceptability judgments for questions with each verb (obtained by averaging ratings obtained from all 60 participants) plotted against their mean negation test scores. The verbs fall into three distinct clusters. On the left-hand side, we have the four factive verbs (*notice*, *know*, *understand*, *complain*), all of which have low negation test scores (i.e., strongly imply the truth of the complement clause). On the far right, we have *think*, which does not presuppose the truth of complement clause, and hence has a high negation test score. The remaining verbs belong to the middle cluster. Notice that while the acceptability ratings for questions with verbs belonging to the middle cluster are generally lower than for *think*, and those for factives are lower still, there is considerable variation in acceptability ratings within clusters. Finally, while *say* and

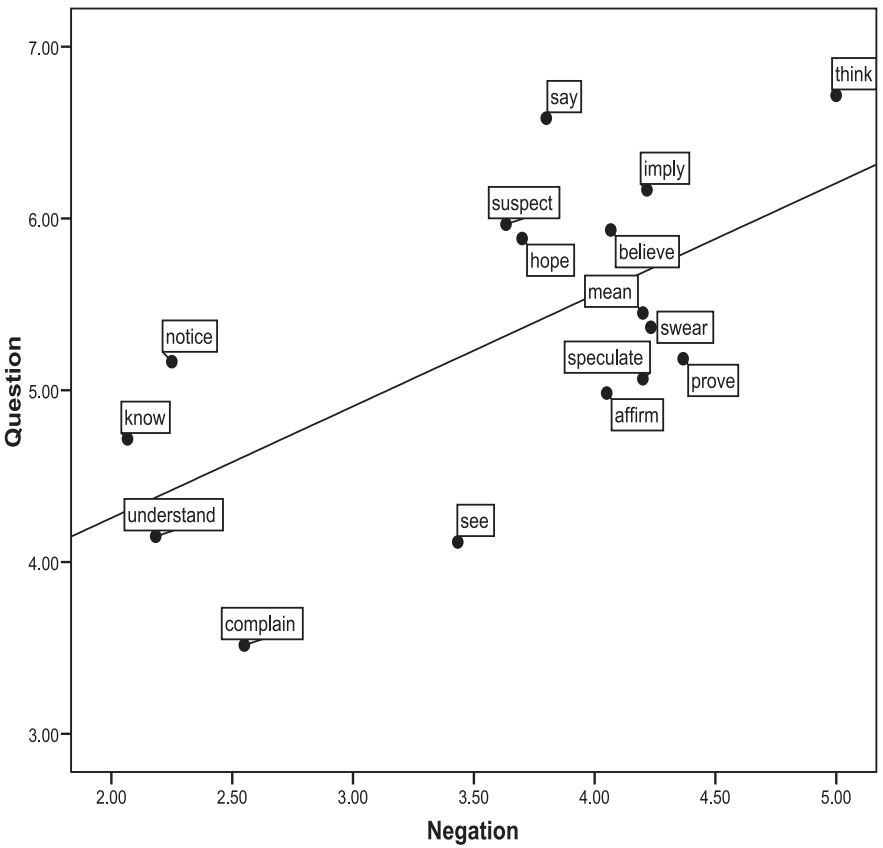

Fig. 4: Relationship between negation test scores and acceptability ratings for LDD questions

think are well above the regression line, they are not obvious outliers. (We will return to this issue later.)

The overall correlation between performance on the two tasks is moderately strong: $r = 0.64$, $p = 0.007$. The correlation between the negation test scores and difference scores, i.e. the measure used by Ambridge and Goldberg, is somewhat stronger: $r = -0.74$, $p = 0.001$ (see Figure 5: note that this correlation is negative because difference scores are calculated by subtracting question ratings from those of the corresponding declaratives). These figures are similar to those obtained by Ambridge and Goldberg (0.58 and −0.83 respectively); in contrast to the results of the latter study, however, the strength of the correlation is similar regardless of whether we use difference scores or raw scores for questions. Thus, the results of this study are consistent with BCI and provide further support for the hypothesis.

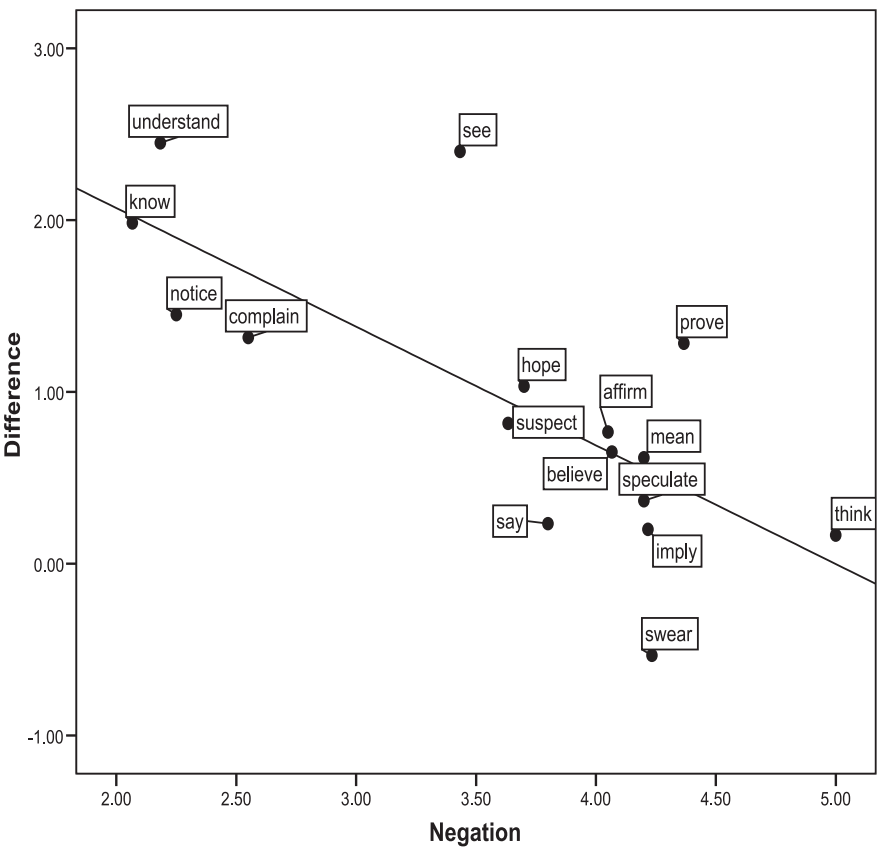

Fig. 5: Relationship between negation test scores and difference scores (dispreference for question scores)

It had been suggested in the introductory section that the use of difference scores in the Ambridge and Goldberg study may have masked lexical effects in LDD questions. A comparison of Figures 4 and 5 confirms that this is a legitimate concern: as in the Ambridge and Goldberg study, the verb *think* is almost on the regression line if we use difference scores, and well above it if we use acceptability ratings for questions. Since this paper is intended to assess the role of both factors, the latter measure will be used in all further analyses.

A somewhat different picture emerges when we look at individual performance. Table 3 shows the correlation coefficients for performance on the two

Table 3: Individual correlations between negation test scores and question scores

Participant	*r*	*p*	Participant	*r*	*p*
1	.713	.002	7	.556	.025
2	.700	.003	8	.519	.039
3	.664	.005	9	.513	.042
4	.637	.008	10	.472	.065
5	.620	.010	11	.467	.068
6	.590	.016	12	.431	.096
13	.418	.107	25	.299	.260
14	.418	.107	26	.286	.284
15	.392	.133	27	.284	.287
16	.386	.140	28	.280	.294
17	.378	.149	29	.261	.329
18	.367	.162	30	.260	.331
19	.364	.165	31	.252	.347
20	.364	.166	32	.248	.355
21	.328	.215	33	.244	.362
22	.327	.217	34	.240	.371
23	.327	.216	35	.237	.378
24	.300	.259	36	.187	.489
37	.151	.576	49	−.036	.895
38	.145	.591	50	−.044	.873
39	.087	.748	51	−.056	.835
40	.069	.800	52	−.065	.811
41	.061	.823	53	−.085	.755
42	.042	.878	54	−.092	.734
43	.036	.893	55	−.170	.529
44	.017	.952	56	−.220	.413
45	.011	.968	57	−.266	.320
46	.004	.988	58	−.282	.289
47	−.009	.973	59	−.468	.067
48	−.035	.899	60	−.493	.052

tasks and their associated probability values computed for each participant separately. To facilitate analysis, the correlation coefficients have been arranged from highest to lowest. As we can see from these data, individual correlation coefficients range from 0.713 – i.e. slightly higher than the group coefficient – to –0.493. Surprisingly, only nine out of the 60 individual coefficients are statistically significant; another three approach significance (p values between 0.05 and 0.10; note that the p values given in the table have *not* been corrected for multiple comparisons). Twenty-four participants have correlation coefficients between 0.42 to 0.19, with the corresponding p values ranging from 0.107 to 0.489. The remaining twenty-four participants have correlation coefficients lower than 0.20; fourteen have negative values. The mean of all individual correlation coefficients is 0.21.

How can we explain this apparent discrepancy between individual and group results? There are two (mutually nonexclusive) possibilities:

1. Individual data are very noisy: thus, we can only obtain reliable estimates of the acceptability of a particular verb in the LDD question construction and the degree to which it backgrounds the complement clause by averaging over judgments obtained from a large number of participants (as Ambridge and Goldberg have done).
2. There are individual differences between speakers: some show evidence of sensitivity to BCI and some do not.

To assess the degree to which either or both of these alternatives can account for the discrepancy, it will be helpful to divide the participants into three groups: "sensitive", "insensitive", and "middle". The "sensitive" group comprises individuals whose correlation coefficients show at least a moderately strong relationship ($r \geq 0.40$) between the two variables which is either significant or approaches significance ($p \leq 0.10$), i.e., participants 1–12. The "insensitive" group comprises individuals whose correlations coefficients are either close to zero with corresponding p values larger than 0.50, or whose correlation coefficients are negative, i.e. participants 37–60. The middle group includes the remaining participants. This division is admittedly somewhat arbitrary. The idea is to have a relatively homogeneous "sensitive" group (hence the cutoff point at $p = 0.10$), and a group which shows no evidence of sensitivity to BCI. Note that the middle and insensitive group are relatively large (24 participants each), which will be useful from the point of view of statistical analysis, since one may expect the data from these participants to be more noisy than from the first group.

We can now compute group correlations between performance on the negation task and acceptability judgment task (i.e., correlations between each group's acceptability rating for each verb, averaged across participants, and the group's negation task ratings, also averaged across participants). For the sensitive group,

$r = 0.86$, $p < 0.001$, while the average of individual correlation coefficients is 0.57. For the middle group, $r = 0.71$, $p = 0.002$, while the individual average is 0.31. Finally, for the insensitive group, $r = 0.20$, $p = 0.461$, and the average of individual correlation coefficients is −0.07. Thus in all three cases, the group correlation coefficients are considerably higher than the mean of individual coefficients of the group members, which suggests that individual data are indeed noisy. However, there are also clear differences between groups, and while one could reasonably argue that participants belonging to the middle group are in fact sensitive to the relationship between backgrounding of the subordinate clause and the acceptability of LDD questions (their low individual correlation coefficients being due simply to noise in individual data), this interpretation is not possible for the last group, which shows no evidence of sensitivity to BCI.

Of course the last group's failure to demonstrate the predicted relationship in an experimental setting does not necessarily mean that individuals belonging to this group have not internalized BCI: it is also possible that they have misunderstood the experimental task, were not paying attention, or were simply uncooperative. This, however, is unlikely, given the high inter-group correlations on all the experimental items. As shown in Table 4, these range from 0.81 to 0.93 for acceptability ratings for declarative sentences, from 0.84 to 0.91 for acceptability ratings for questions, and from 0.75 to 0.95 for the negation task, and all are significant at the 0.001 level or below. In other words, all three groups agree in their judgments of acceptability of LDD questions and declaratives with different verbs and the degree to which a particular verb presupposes the truth of the complement clause. It is thus unlikely that the fact that the insensitive group did not show a relationship between the two variables simply because they failed to engage with the task: the main difference between the three groups appears to be in individual sensitivity to BCI.

Table 4: Correlations between the three groups' judgments

Groups	Declarative	Question	Negation
Sensitive and middle	0.91	0.89	0.95
Middle and insensitive	0.93	0.91	0.83
Sensitive and insensitive	0.81	0.84	0.75

3.3 Testing the LTH: Lexical effects

According to the Lexical Template Hypothesis, speakers store partially specific templates for LDD questions (*WH do you think S-GAP?*, *WH did you say S-GAP?*) and

produce "prototypical" LDD questions simply by inserting lexical material into the slots, while unprototypical questions also require modifying the template. The hypothesis thus predicts that LDD questions with *think* and *say* will be judged better than questions with other verbs. This prediction was confirmed in the study described in Dąbrowska (2008) and also in the present study: for *think* v. other verbs, $t(59) = 15.66$, $p < 0.001$; for *say* v. other verbs: $t(59) = 11.31$, $p < 0.001$. In fact, when there was no overt complementizer, questions with *think* and *say* were judged to be virtually perfect, receiving mean ratings of 6.90 and 6.87 respectively.

However, BCI also predicts that LDD questions with *think* and *say* will be rated more acceptable than questions with other verbs, since these verbs do not presuppose the truth of the complement clause. The question, then, is whether interrogatives with *think* and *say* are judged to be significantly better than predicted by the regression equation computed on the basis of the other verbs. As can be seen

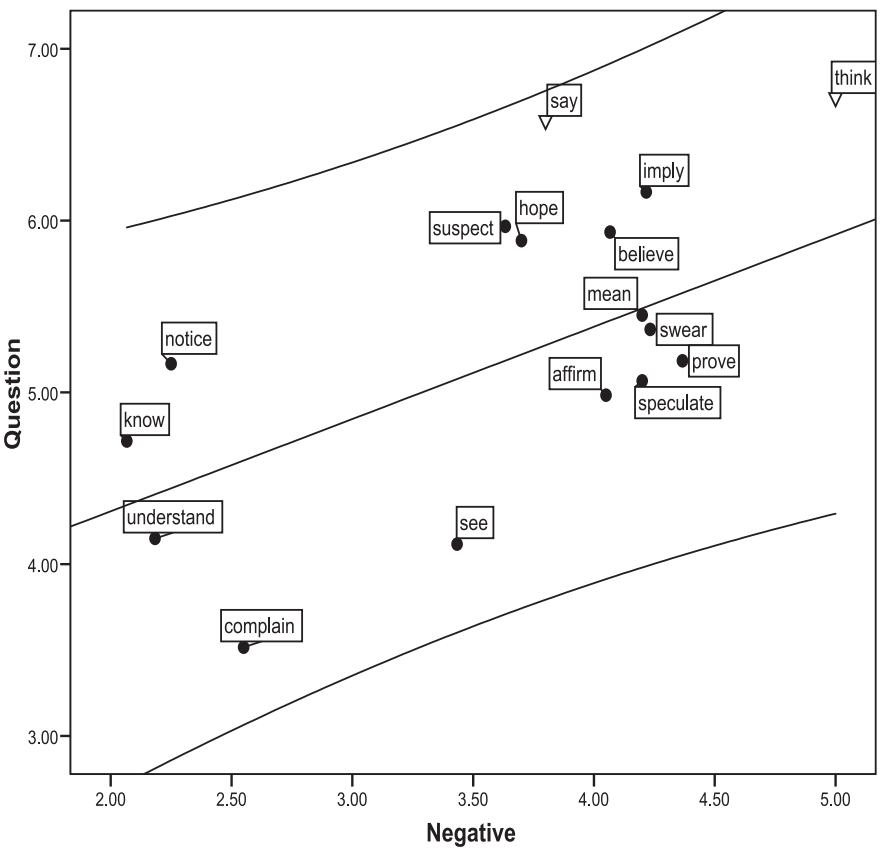

Fig. 6: Actual values for *think* and *say* plotted against the regression line for the other 14 verbs

from Figure 6, although the values for both verbs are well above the regression lines, they are within the 95% confidence intervals for individual values. Note, however, that the intervals are very wide (since we have only a small number of data points), and the in order to fall outside the intervals, the acceptability ratings for questions with both verbs would have to be above 7 (i.e., above the maximum value on the scale). Thus, the data for the entire sample are inconclusive.

A clearer picture emerges when we compare the performance of the three groups identified in the section on the role of the verb. Figure 7 shows the results for the "sensitive" group, i.e. participants whose individual correlation coefficients either reached or approached statistical significance. As we can see in the figure, the regression line is quite steep and the confidence intervals fairly narrow. *Say* and *think* are both very close to the regression line, the former falling

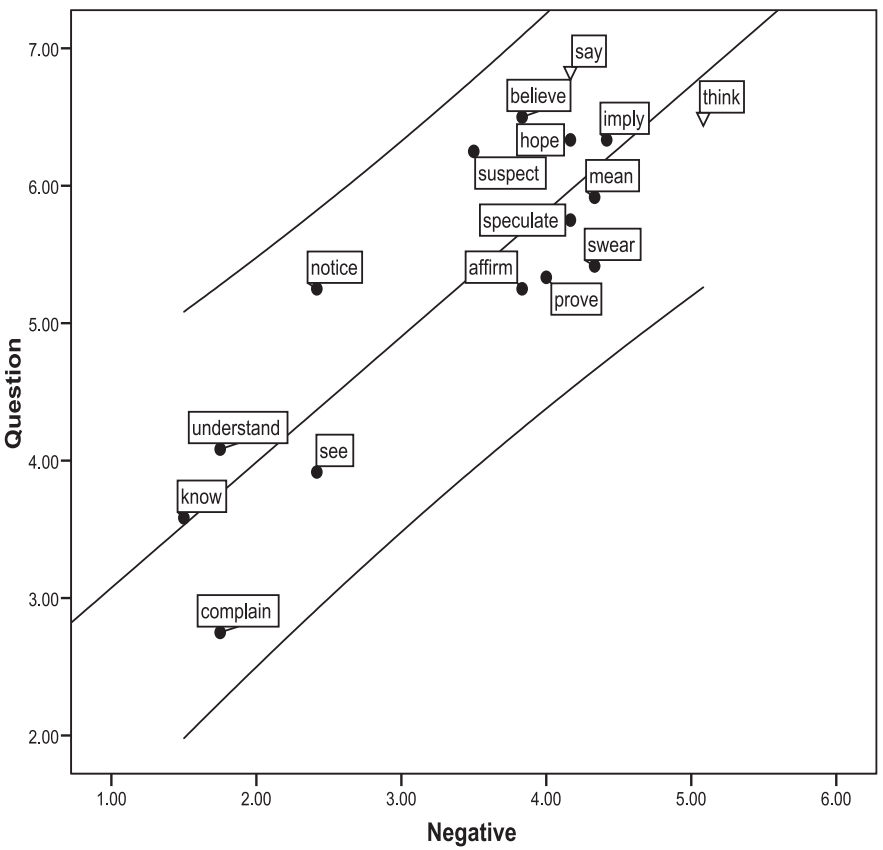

Fig. 7: Actual values for *think* and *say* plotted against the regression line for the other 14 verbs: "Sensitive" group

just above it and the latter just below (which is presumably a ceiling effect). In other words, this group behave exactly as predicted by BCI.

The "insensitive" group, on the other hand, is very different (see Figure 8). The regression line almost flat, indicating virtual lack of relationship between the two variables, and both verbs are well above it, with *think* slightly above the upper 95% confidence interval, and *say* just below it. This is quite remarkable, given the width of the intervals. Thus, the insensitive group behave as predicted by LTH: they show no sensitivity to the relationship between the degree to which a particular verb backgrounds the subordinate clause and acceptability of an LDD question with that verb, and appear to base their acceptability judgments purely on frequency. A closer inspection of Figure 9 also helps us to understand how the "insensitive" participants were able to give ratings for both tasks that correlated

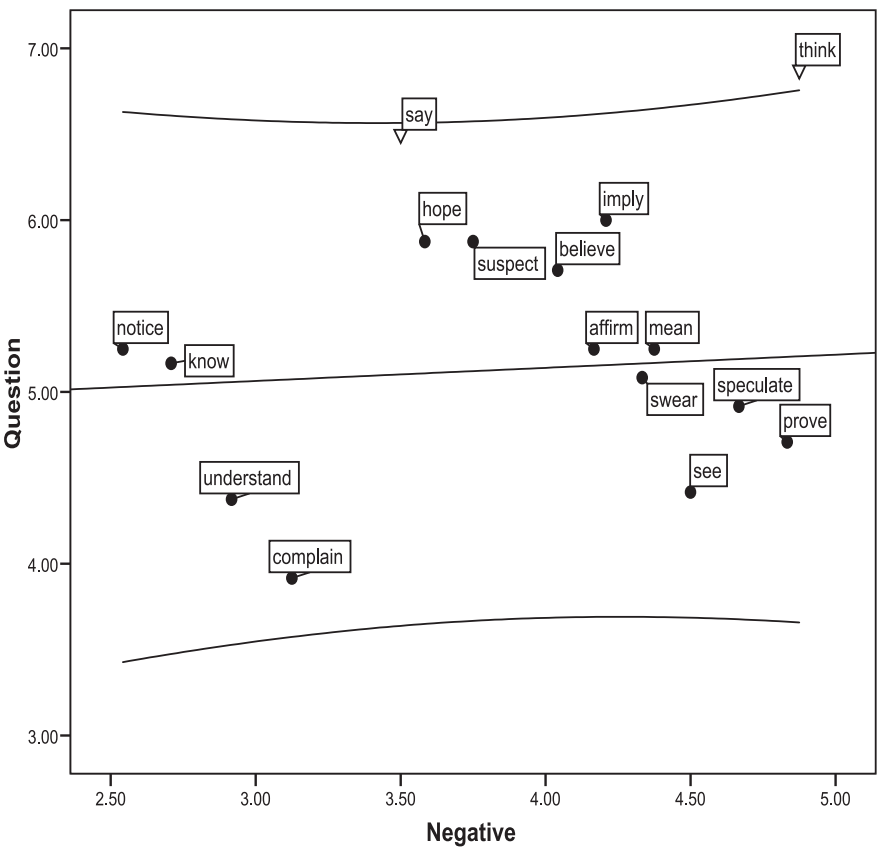

Fig. 8: Actual values for *think* and *say* plotted against the regression line for the other 14 verbs: "Insensitive" group

highly with those provided by the other groups. This is due largely to the fact that on the negation test, they consistently gave low ratings to sentences with factive verbs (*notice*, *know*, *understand*, *complain*), while on the acceptability judgment test, they consistently gave very high ratings to questions with *think* and *say*[1] – clearly facts about language that can be learned independently of each other.

The middle group's behaviour, as one might expect, falls between the two extremes: *think* and *say* are well above the regression line but within 95% confidence intervals for individual values (Figure 9). There are several possible ex-

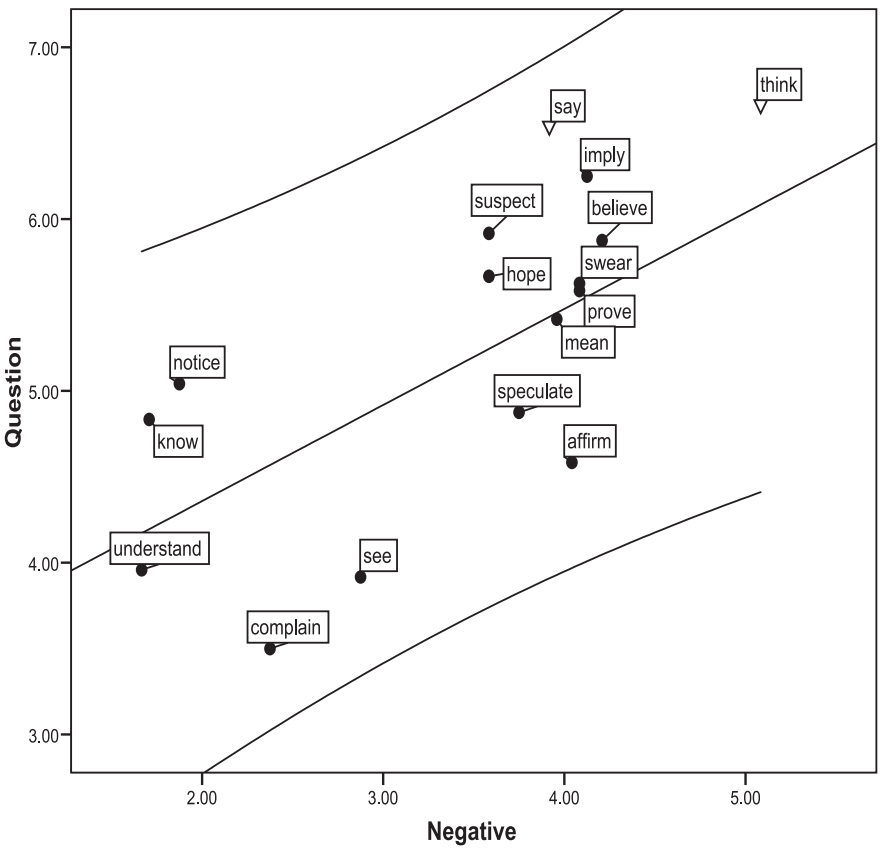

Fig. 9: Actual values for *think* and *say* plotted against the regression line for the other 14 verbs: "Middle" group

1 There are also other similarities between the groups, e.g. all three groups appear to give particularly low ratings to questions with *complain* and *see*. This is a lexical effect which is not predicted by either BCI or the LTH.

planations for this pattern of results. The group may simply be heterogeneous, i.e. contain some individuals who are sensitive to BCI and some who are not, or it may contain individuals who show only a weak sensitivity to the principle. A more intriguing possibility is that they are like the third group in that they are not sensitive to BCI and have simply learned about the meanings of the verbs and about the strength of the relationship between individual complement-taking verbs and the LDD construction independently of each other, but are more sensitive to differences between verbs than the third group. This would enable them to approximate the behaviour of the first group without actually sharing their competence. In other words, on this interpretation, we have three groups of speakers: those who "know" BCI (implicitly, or course, rather than explicitly); those who don't know it, but behave as if they did; and those who don't know it, and approximate the behaviour of the first group only at a very gross level, i.e., accept LDD questions with *say* and *think* much more readily than questions with other complement-taking verbs, but show no sensitivity to differences within the "other" group.

Clearly further research is necessary to determine exactly what the status of BCI is for individual speakers. It may be worth pointing out, however, that the last possibility, though clearly speculative, is not as far-fetched as it might initially appear. Dąbrowska (2008) describes an experimental study of the Polish genitive inflection which suggests that this state of affairs can indeed exist in a language. The Polish genitive masculine has two endings, *-a* and *-u*, but no rule determining the choice of ending. There are, however, some fairly reliable statistical tendencies: for instance, masculine nouns designating small, easily manipulable objects usually take *-a*, while masculine nouns designating substances usually take *-u*. In the experiment described by Dąbrowska, adult native speakers of Polish were asked to supply the genitive masculine form of nonce nouns which referred either to objects or to substances. The results showed that as a group, the speakers were sensitive to the contrast, i.e. they were more likely to use *-u* with nonce nouns designating substances. However, a closer analysis revealed that the group data masked individual differences. Two of the twenty individuals tested performed at ceiling (i.e., consistently chose *-u* with nonce nouns designating substances and *-a* with nonce nouns designating objects), while the remaining 18 performed at chance, which suggests that they had not learned the relevant generalization. What is interesting is that all Polish speakers appear to consistently use *-u* with most real nouns denoting substances and *-a* with real nouns denoting objects: in other words, there is a regularity in the language which is not represented in the minds of most of its speakers. One would expect such a situation to be inherently unstable; yet it has apparently persisted for several centuries (Klemensiewicz et al. 1955), and there are no signs of it disappearing. In fact, when

new borrowings enter the language, they tend to acquire the endings consistent with their meaning: e.g. *teflon, popcorn, tweed, nylon* take *-u*, while *hot dog, walkman, jeep, skaner, marker* take *-a*. It seems that consistent use by a small minority is sufficient to fix the pattern in the language even if the majority of speakers have not learned the relevant generalization and simply learn the endings on a noun-by-noun basis. This enables them to behave *as if* they knew the pattern when in fact they do not, since they are not able to generalize it to novel nouns.

3.4 Complementizer effects

The acceptability judgment data were analysed by means of two 2 (construction) × 2 (complementizer) ANOVAs, one by participant and one by item. The analysis by participant revealed a main effect of construction, $F(1,59) = 215.77$, $p < 0.001$, η_p^2 (partial eta squared) = 0.785, showing that acceptability ratings for LDD questions were significantly lower than those for declaratives. The main effect of complementizer was not significant, but there was a construction × complementizer interaction: $F(1,59) = 8.106$, $p = 0.006$, $\eta_p^2 = 0.121$ (see Figure 10). Further analysis by means of paired-samples *t*-tests showed that questions were judged better when they contained a zero complementizer: $t(59) = 2.00$, $p = 0.05$, $d = 0.30$; the difference between declaratives with and without *that* was not statistically significant. ANOVA by item showed analogous results: there was a main effect of construction, $F(1,15) = 20.82$, $p < 0.001$, $\eta_p^2 = 0.581$, and a construction × complementizer interaction, $F(1,15) = 6.51$, $p = 0.022$, $\eta_p^2 = 0.303$. These results replicate those reported in Dąbrowska 2008.

Can BCI explain this interaction? It has been suggested (see e.g. Kearns 2007; Verhagen 2005) that the use of the zero complementizer foregrounds the subor-

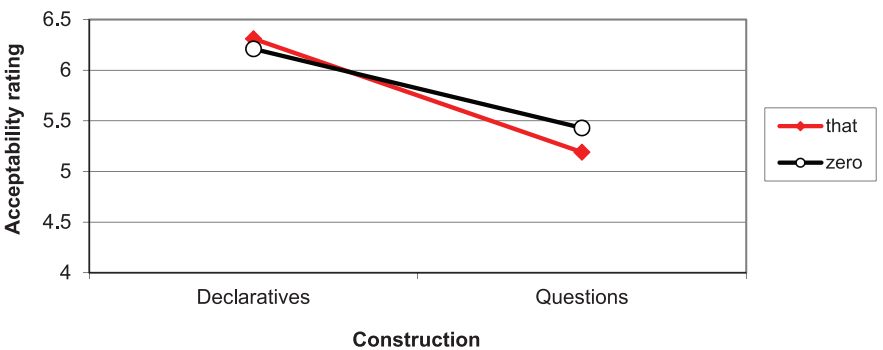

Fig. 10: The construction × complementizer interaction

Table 5: Acceptability ratings and negation test results for sentences with and without *that*

	that		zero	
	Mean	***SD***	**Mean**	***SD***
Declaratives	6.31	0.51	6.21	0.58
Questions	5.19	0.87	5.43	0.77
Negation test	3.55	0.91	3.70	1.01

dinate clause. Thus BCI predicts that subordinate clauses with the zero complementizer should have higher negation test scores than subordinate clauses introduced by *that*. As we can see from Table 5, this is indeed the case, although the difference is very small and not statistically significant: in the analysis by participant, $t(59) = 1.05$, $p = 0.299$, $d = 0.16$; in the analysis by item, $t(15) = 1.77$, $p = 0.098$, $d = 0.17$. These results suggest that the higher acceptability ratings for questions with the zero complementizer are not attributable to its foregrounding effects.

Group data, however, can mask individual differences: that is to say, it is possible that some individuals are sensitive to the contrast between sentences with and without *that*, even if the *t*-test result for the group as a whole is not significant. If this were the case, BCI predicts that these individuals should also show a stronger dispreference for LDD questions with *that*.

In order to examine the extent of individual differences in sensitivity to the backgrounding properties of *that*, mean negation test scores for sentences with and without an overt complementizer were computed for each participant separately and compared by means of *t*-tests. The results are presented in Table 6. As shown in the table, ten participants show evidence of sensitivity to the foregrounding properties of the zero complementizer, i.e. they gave higher negation test scores to sentences with no overt complementizer and the difference is either statistically significant or approaches significance ($p \leq 0.10$). Another 21 participants gave higher scores to sentences without *that*, but the difference is not statistically significant. The remaining 29 participants show no difference or a difference in the opposite direction; for six of them, the difference was statistically significant or approached significance.

In order to determine whether the individuals who are more sensitive to the backgrounding properties of the complementizer show a stronger dispreference for LDD questions with *that*, the participants were divided into three groups, as described in the preceding paragraph, and a 2 (construction) × 2 (complementizer) × 3 (group) ANOVA was carried out. The analysis revealed a main effect of construction and a construction × complementizer interaction, but no interactions between group and any of the other variables: all three groups showed

Table 6: Individual sensitivity to the foregrounding properties of the zero complementizer

Participant	*that* difference	t(14)	p	Participant	*that* difference	t(14)	p
40	2.750	3.274	.006	1	1.500	2.256	.041
12	2.375	2.747	.016	50	1.625	2.089	.055
51	2.500	2.736	.016	6	1.500	1.984	.067
39	1.750	2.701	.017	23	1.250	1.776	.098
33	1.375	2.280	.039	41	.875	1.758	.101
34	1.125	1.655	.120	15	.375	.798	.438
26	1.250	1.587	.135	54	.250	.683	.506
24	1.125	1.580	.136	2	.625	.574	.575
19	1.500	1.542	.145	35	.750	.546	.594
58	.750	1.528	.149	20	.375	.505	.621
7	.750	1.461	.166	29	.375	.424	.678
4	.750	1.232	.238	18	.375	.377	.712
57	.625	1.077	.300	53	.375	.346	.735
52	.875	1.043	.314	45	.375	.306	.764
9	.375	.851	.409	46	.125	.290	.776
27	.375	.814	.429				
17	.000	.000	1.000	14	−.625	−.755	.463
30	.000	.000	1.000	43	−.750	−.839	.416
48	.000	.000	1.000	38	−.750	−1.033	.319
59	−.125	−.193	.849	60	−.750	−1.128	.278
37	−.250	−.243	.812	10	−.875	−1.133	.276
42	−.250	−.247	.809	36	−1.125	−1.142	.273
16	−.125	−.277	.786	28	−.625	−1.330	.205
5	−.250	−.290	.776	47	−1.375	−1.511	.153
3	−.250	−.317	.756	31	−1.500	−1.845	.086
44	−.375	−.431	.673	49	−1.625	−1.914	.076
13	−.250	−.457	.655	11	−1.875	−2.295	.038
8	−.375	−.509	.619	25	−2.625	−2.411	.030
32	−.500	−.564	.582	22	−2.000	−3.121	.008
55	−.250	−.672	.513	56	−2.000	−3.347	.005
21	−.375	−.693	.499				

Note: That difference is the difference between a participant's negation test scores for sentences with and without an overt complementizer. Thus a positive figure in this column (and a positive *t*-test value) indicates that the participant gave higher scores to sentences with the zero complementizer.

a similar dispreference for *that* in questions as compared to declaratives. These results suggest that BCI cannot explain complementizer effects.

4 Conclusion

The experiment described in this paper revealed a moderately strong and highly significant correlation between the degree to which a particular verb backgrounds the subordinate clause and the acceptability of LDD questions with that verb. This replicates the results obtained by Ambridge and Goldberg (2006), and provides further support for their claim that backgrounded constituents are islands to extraction. However, an analysis of individual data presents a somewhat different picture. As we have seen, only about 20% of the informants showed evidence of sensitivity to the constraint. 40% were apparently not sensitive, in that their individual correlations were either close to zero or negative, and their group correlation was also not significant. The status of BCI in the remaining 40% of informants is uncertain. Crucially, however, the performance of the "insensitive" and "middle" groups on both tasks was highly correlated with that of the "sensitive" group. This suggests that the "insensitive", and possibly also the "middle" group apparently learned about meanings of the verbs and the frequency with which they occur in LDD question construction independently of each other and were able to approximate the behaviour of the sensitive group without sharing the same knowledge. In other words, it appears that, while BCI is a valid generalization about the English language, it is not necessarily a generalization that is captured in the mental grammars of all, or even most, speakers of English.

Regardless of whether they were sensitive to BCI or not, all three groups gave very high ratings to LDD questions with *think* and *say*. For the sensitive and middle group, these ratings fell within confidence intervals for individual values predicted by the equation derived from data for the other verbs. However, to fall outside the intervals, the values would have to be above the maximum value on the scale used in the experiment, so we cannot conclude very much from this fact. In the insensitive group, acceptability ratings for questions with *think* and *say* were significantly higher than predicted by BCI, which shows that speakers belonging to this group at least rely on lexical templates.[2]

2 It should be pointed out that it does not necessarily follow that these speakers don't *also* have a more general representation in addition to the templates. However, since it is virtually impossible to demonstrate that speakers do *not* have a particular representation, the burden of proof rests with those who maintain that they do.

The results reported above also suggest that BCI cannot account for the observed differences between questions with and without an overt complementizer: the experiment did not reveal a significant difference in negation test scores for sentences with and without *that,* although there was a trend in the predicted direction. It is possible, of course, that the trend would turn out to be statistically significant in a larger sample; however, given the size of the effect, it is unlikely to account for the observed differences in acceptability judgments.

Why then are LDD questions without an overt complementizer judged more acceptable than questions with *that*? As noted earlier, *think* and *say* occur particularly frequently in the main verb position in LDD questions: in fact, 86% of all LDD questions in the British National Corpus have either *think* or *say* as the main clause verb.[3] As shown by Ambridge and Goldberg, and corroborated by the present study, this can be explained by the fact that their meanings are particularly compatible with the LDD question construction in that they do not presuppose the truth of the embedded clause that they introduce. Both of these verbs also strongly favour the zero complementizer (Kearns 2007, Roland, Dick and Elman 2007). Thus, the LDD templates (*WH do you think S-GAP? WH did you say S-GAP?*) do not contain a complementizer; and if speakers produce LDD questions with other verbs by modifying the template, as the LTH proposes, questions without *that* are preferred even with other verbs.

Thus the experiment described in this paper provides evidence supporting both hypotheses. It is also clear that neither can account for all of the data: LTH cannot explain why the LDD question construction tends to prefer some verbs over others; and BCI cannot account for complementizer effects or the outlier status of questions with *think* and *say* for some speakers. This is not a problem, however: as pointed out in the introduction, BCI and the Lexical Template Hypothesis are not necessarily incompatible with each other, and could be seen as capturing different aspects of the same phenomenon. It is clear that BCI – or other functional constraints, for that matter – does not shape mental grammars directly: BCI shapes usage which in turn shapes grammars. In other words, BCI explains why speakers produce certain combinations of words frequently and avoid others, but to understand what happens later – why there are differences between languages and language varieties – we need to examine the social and psychological mechanisms that govern usage. It is well established that frequently recurring patterns often acquire unit status (Langacker 2000; Bybee 2006); this makes

3 These two verbs are also quite frequent in declaratives, but their relative frequency in the latter construction is considerably lower: only 56% of declaratives with finite verb complement clauses in the BNC have either *think* or *say* as the matrix verb, as compared with 86% of LDD questions.

them easier to use, which results in even greater frequency, and, in some cases, leads to the emergence of new properties, such as the strong preference for LDD questions without an overt complementizer observed in this and earlier studies.

References

Ambridge, Ben & Adele E. Goldberg. 2008. The island status of clausal complements: evidence in favor of an information structure explanation. *Cognitive Linguistics* 19. 357–389.

British National Corpus, The, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: <http://www.natcorp.ox.ac.uk/>.

Bybee, Joan. 2006. From usage to grammar: the mind's response to repetition. *Language* 82. 711–733.

Dąbrowska, Ewa. 2004. *Language, Mind and Brain. Some Psychological and Neurological Constraints on Theories of Grammar*. Edinburgh: Edinburgh University Press.

Dąbrowska, Ewa. 2008. Questions with 'unbounded' dependencies: A usage-based perspective. *Cognitive Linguistics* 19. 391–425.

Dąbrowska, Ewa. 2010. Naive v. expert competence: An empirical study of speaker intuitions. *The Linguistic Review* 27. 1–23.

Dąbrowska, Ewa, Caroline Rowland & Anna Theakston. 2009. The acquisition of questions with long-distance dependencies. *Cognitive Linguistics* 20. 571–596.

Frazier, Lyn & Charles Clifton, Jr. 1989. Successive cyclicity in the grammar and the parser. *Language and Cognitive Processes* 4. 93–126.

Goldberg, Adele E. (2006). *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.

Hawkins, J. A. (1999). Processing complexity and filler-gap dependencies across grammars. *Language, 75*, 244–285.

Hawkins, John A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Kearns, Kate. 2007. Epistemic verbs and zero complementizer. *English Language and Linguistics* 11. 475–505.

Klemensiewicz, Zenon, Tadeusz Lehr-Spławiński & Sranisław Urbańczyk. 1955. *Gramatyka historyczna języka polskiego*. Warszawa: PWN.

Kluender, Robert & Marra Kutas. 1993. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8(4). 573–633.

Langacker, Ronald W. 2000. A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-Based Models of Language*, 1–63. Stanford, CA: CSLI Publications.

Roland, Douglas, Frederic Dick & Jeffery Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language* 57, 348–379.

Verhagen, Arie. 2005. *Constructions of Intersubjectivity: Discourse, Syntax and Cognition*. Oxford: Oxford University Press.

Verhagen, Arie. 2006. On subjectivity and 'long distance *Wh*-movement'. In A. Athanasiadou, C. Canakis & B. Cornillie (eds.), *Various Paths to Subjectivity*, 323–346). Berlin/New York: Mouton de Gruyter.

Appendix: Questionnaire used in the study (version 1A)

Part 1

Please rate each of the sentences below for how acceptable you find it, on a scale from 7 (completely acceptable) to 1 (completely unacceptable). Write your answer in the blank next to the sentence.

Please judge the sentences only on how acceptable you find them (and not, for example, whether the event they describe is plausible or implausible, good or bad etc.). Acceptability is a sliding scale and not a yes/no judgment: people tend to differ in their judgments of how acceptable sentences are.

Remember:

7 = Perfect (completely acceptable)
1 = Terrible (completely unacceptable)

1. _____ I noticed that Mike drinks vodka.
2. _____ What did you put forward the hypothesis that the scientist discovered?
3. _____ I complained that Chris likes seafood.
4. _____ What did you say Laura eats?
5. _____ I think that Kate likes Italian food.
6. _____ What did you speculate Andrew wants?
7. _____ I believe Sue plays bridge.
8. _____ What did you affirm that Mike plays?
9. _____ The manager not implied you knew about it.
10. _____ What did you swear that Liz drives?
11. _____ His cousin doesn't thinks we lied.
12. _____ What did you notice that Anne drinks?
13. _____ I hope Paul drinks gin.
14. _____ What do you know that Bob needs?
15. _____ What did you discover the fact that the criminals stole?
16. _____ What did you complain that Neil likes?
17. _____ What do you think that got lost?
18. _____ I mean Laura drives a Porsche.
19. _____ Your brother doesn't believes the man is telling the truth.
20. _____ I swore that Neil drives a Mercedes.
21. _____ What do you mean Sue drives?

22. _____ The mother doesn't knows Julia was absent.
23. _____ I see Eve wants an ice-cream.
24. _____ What did Claire make the claim that she read?
25. _____ What do you think that Chris likes?
26. _____ What did Paul hear the rumour that I found?
27. _____ I understand that Bob eats meat.
28. _____ Her husband not claimed they left.
29. _____ What do you believe Paul plays?
30. _____ What did you say that will kill cockroaches?
31. _____ I proved Andrew reads *The Daily Mail*.
32. _____ What do you suspect that Lucy reads?
33. _____ I know that Lucy needs help.
34. _____ What do you believe that will turn up in the evening?
35. _____ I speculated Anne wants a dessert.
36. _____ The teacher not suspected she remembered.
37. _____ I suspect that Jack reads *The Times*.
38. _____ What do you understand that Kate eats?
39. _____ What did you guess that exploded?
40. _____ I said Ian eats fish.
41. _____ The girl doesn't remembers Peter borrowed this.
42. _____ What did you imply Ian needs?
43. _____ Your sister not believed I forgot.
44. _____ What did you prove Julie reads?
45. _____ I implied Julie needs money.
46. _____ What do you hope Eve drinks?
47. _____ I affirmed that Liz plays poker.
48. _____ What do you see Jack wants?

Part 2

Here, you will be given two statements. Your task is to decide the extent to which the first statement implies the second statement, again using a scale from 1 to 7.

Consider the example sentence pairs in A–C below:

(A) *Bob left early.*
Bob didn't leave early. ____1____

The first statement strongly implies that the second statement is NOT true, so in this case you should choose 1, as shown above.

(B) *Bob left the party early.*
Bob left the party. ____7____

This time, the first statement strongly implies that the second statement IS true, so this time, you should choose 7 as shown above.

(C) *Bob might leave the party late.*
Bob left the party early. ____4____

This time, the first statement neither implies nor does not imply the second statement, so here you would choose 4 as shown above.

We are interested in what ordinary people typically imply with their everyday statements. Bearing these examples in mind, please rate the pairs below for the extent to which the first statement implies that the second statement is true. That is, if you heard a person say [Statement 1], to what extent would you assume that they are implying [Statement 2].

Remember:

7 = second statement definitely TRUE
1 = second statement definitely NOT TRUE.

1. Liz doesn't mean Laura drives a Porsche.
   Laura doesn't drive a Porsche. _____
2. Laura doesn't understand that Bob eats meat.
   Bob doesn't eat meat. _____
3. Sue doesn't think that Kate likes Italian food.
   Kate doesn't like Italian food. _____
4. Julie doesn't suspect that Jack reads *The Times*.
   Jack doesn't read *The Times* ._____
5. Anne doesn't hope Paul drinks gin.
   Paul doesn't drink gin. _____
6. Jack didn't affirm that Liz plays poker.
   Liz doesn't play poker. _____
7. Chris didn't imply Julie needs money.
   Julie doesn't need money. _____
8. Andrew doesn't see Eve wants an ice-cream.
   Eve doesn't want an ice-cream. _____

9. Neil didn't say Ian eats fish.
   Ian doesn't eat fish. _____
10. Lucy didn't notice that Mike drinks vodka.
    Mike doesn't drink vodka. _____
11. Ian doesn't know that Lucy needs help.
    Lucy doesn't need help. _____
12. Kate didn't prove Andrew reads *The Daily Mail*.
    Andrew doesn't read *The Daily Mail*. _____
13. Bob didn't speculate Anne wants a dessert.
    Anne doesn't want a dessert. _____
14. Eve didn't swear that Neil drives a Mercedes.
    Neil doesn't drive a Mercedes. _____
15. Mike doesn't believe Sue plays bridge.
    Sue doesn't play bridge. _____
16. Paul didn't complain that Chris likes seafood.
    Chris doesn't like seafood. _____

Finally . . . . some information about you:

Age ____ Gender (M/F) ____

Are you a native speaker of English? (Y/N) _____

This is all. Thank you very much for completing the questionnaire!